Inferences from Prior-based Loss Functions

Michael Evans
Department of Statistics
University of Toronto
Toronto, ON M5S 3G3
mevans@utstat.utoronto.ca

Gun Ho Jang
Department of Biostatistics and Epidemiology
University of Pennsylvania
Philadelphia, PA 19104, USA
gunjang@upenn.edu

Abstract

Inferences that arise from loss functions determined by the prior are considered and it is shown that these lead to limiting Bayes rules that are closely connected with likelihood. The procedures obtained via these loss functions are invariant under reparameterizations and are Bayesian unbiased or limits of Bayesian unbiased inferences. These inferences serve as well-supported alternatives to MAP-based inferences.

Key words and phrases: loss functions, relative surprise, lowest posterior risk region, Bayesian unbiasedness.

1 Introduction

Suppose we have a sampling model, given by a collection of densities $\{f_\theta : \theta \in \Theta\}$ with respect to a support measure μ on sample space $\mathcal{X}$, and a proper prior, given by density π with respect to support measure ν on Θ. When we observe data x these ingredients lead to the posterior on Θ with density given by $\pi(\theta \,|\, x) = \pi(\theta) f_\theta(x)/m(x)$ with respect to support measure ν where $m(x) = \int_\Theta \pi(\theta) f_\theta(x)\, \nu(d\theta)$.

One can determine inferences based on these ingredients alone. For example, suppose we are interested in a characteristic $\psi = \Psi(\theta)$ where $\Psi : \Theta \to \Psi$ and we let Ψ stand for both the space and mapping to conserve notation. The highest posterior density (hpd), or MAP-based, approach to determining inferences constructs credible regions of the form

$$H_\gamma(x) = \{\psi_0 : \pi_\Psi(\psi_0 \,|\, x) \geq h_\gamma(x)\} \tag{1}$$

where $\pi_\Psi(\cdot \,|\, x)$ is the marginal posterior density with respect to a support measure ν_Ψ on Ψ, and $h_\gamma(x)$ is chosen so that $h_\gamma(x) = \sup\{k : \Pi_\Psi(\{\psi : \pi_\Psi(\psi \,|\, x) \geq k\} \,|\, x) \geq \gamma\}$. It follows from (1) that, if we want to assess the hypothesis $H_0 : \Psi(\theta) = \psi_0$, then we can use the tail probability given by $1 - \inf\{\gamma : \psi_0 \in H_\gamma(x)\}$. Furthermore, the class of sets $H_\gamma(x)$ is naturally "centered" at the posterior mode (when it exists uniquely) as $H_\gamma(x)$ converges to this point as $\gamma \to 0$. The use of the posterior mode as an estimator is commonly referred to as MAP (maximum *a posteriori*) estimation. We can then think of the size of the set $H_\gamma(x)$, say for $\gamma = 0.95$, as a measure of how accurate the MAP estimator is in a given context. Furthermore, we have that when Θ is an open subset of a Euclidean space, then $H_\gamma(x)$ minimizes volume among all γ-credible regions. The use of MAP-based inferences is very common in machine learning contexts, see, for example, Bishop (2006).

It is well-known, however, that hpd inferences suffer from a serious defect. In particular, in the continuous case hpd inferences are not invariant under reparameterizations. For example, this means that if $\psi_{\mathrm{MAP}}(x)$ is the MAP estimate of ψ, then it is not necessarily true that $\Upsilon(\psi_{\mathrm{MAP}}(x))$ is the MAP estimate of $\tau = \Upsilon(\psi)$ when Υ is a 1-1, smooth transformation. The noninvariance of a statistical procedure seems very unnatural as it implies that the statistical analysis depends on the parameterization and typically there does not seem to be a good reason for this.

A class of inferences, similar to hpd inferences, avoids this lack of invariance. These are referred to as *relative surprise inferences* and are based on the regions

$$C_\gamma(x) = \{\psi : \pi_\Psi(\psi \,|\, x)/\pi_\Psi(\psi) \geq c_\gamma(x)\} \tag{2}$$

where π_Ψ is the marginal prior density with respect to a support measure ν_Ψ on Ψ, and $c_\gamma(x) = \sup\{k : \Pi_\Psi(\{\psi : \pi_\Psi(\psi \,|\, x)/\pi_\Psi(\psi) \geq k\} \,|\, x) \geq \gamma\}$. The hypothesis $H_0 : \Psi(\theta) = \psi_0$ is assessed by computing the tail probability

$$1 - \inf\{\gamma : \psi_0 \in C_\gamma(x)\} = \Pi_\Psi(\pi_\Psi(\psi \,|\, x)/\pi_\Psi(\psi) \leq \pi_\Psi(\psi_0 \,|\, x)/\pi_\Psi(\psi_0) \,|\, x). \tag{3}$$

We refer to $\pi_\Psi(\psi \,|\, x)/\pi_\Psi(\psi)$ as the *relative belief ratio* of ψ as it measures how beliefs in ψ being the true value change from *a priori* to *a posteriori*. The relative surprise terminology then comes from (3) as this is measuring how surprising the value ψ_0 is by comparing its relative belief ratio to the relative belief ratios of other values of ψ. The corresponding estimator is given by the maximizer of the ratio $\pi_\Psi(\psi \,|\, x)/\pi_\Psi(\psi)$, which we refer to as the *least relative surprise estimator* (LRSE), and denote as $\psi_{\mathrm{LRSE}}(x)$. Note that $\psi_{\mathrm{LRSE}}(x)$ is the least surprising value as it maximizes (3). Beyond their invariance these inferences have many optimality properties in the class of all Bayesian inferences as documented in Evans (1997), Evans, Guttman and Swartz (2006), Evans and Shakhatreh (2008) and Jang (2010). In this paper we will establish optimal decision-theoretic properties for relative surprise inferences.

The idea of measuring surprise based on how beliefs change from *a priori* to *a posteriori* and using this for inference, has arisen in other discussions. For

example, see Baldi and Itti (2010) for the use and development of this idea in the context of learning.

While hpd and relative surprise inferences may seem quite natural, another ingredient is often added to the formulation of a statistical problem, namely, a loss function. For this we have an action space Ψ, a function $\Psi : \Theta \to \Psi$, such that $\Psi(\theta)$ is the correct action when θ is true, and a loss function $L : \Theta \times \Psi \to [0, \infty)$ satisfying $L(\theta, \Psi(\theta)) = 0$, i.e., there is no loss when we take the correct action. The goal of a statistical decision analysis is then to find a decision function $\delta : \mathcal{X} \to \Psi$ that minimizes the prior risk $r(\delta) = \int_\Theta \int_\mathcal{X} L(\theta, \delta(x)) f_\theta(x) \pi(\theta)\, \mu(dx)\, \nu(d\theta) = \int_\mathcal{X} r(\delta \,|\, x) m(x)\, \mu(dx)$ where $r(\delta \,|\, x) = \int_\Theta L(\theta, \delta(x)) \pi(\theta \,|\, x)\, \nu(d\theta)$ is the posterior risk. Such a δ is called a Bayes rule and clearly a δ that minimizes $r(\delta \,|\, x)$ for each x is a Bayes rule. Further discussion of decision theory can be found in Berger (1985).

As noted in Bernardo (2005) a decision formulation also leads to credible regions for ψ, namely, a γ-*lowest posterior loss credible region* is defined by

$$L_\gamma(x) = \{\psi : r(\psi \,|\, x) \le l_\gamma(x)\} \tag{4}$$

where $l_\gamma(x) = \inf\{k : \int_{\{\psi_0 : r(\psi_0 \,|\, x) \le k\}} \pi_\Psi(\psi \,|\, x)\, \nu_\Psi(d\psi) \ge \gamma\}$. Note that ψ in (4) is interpreted as the decision function that takes the value ψ constantly in x. Clearly as $\gamma \to 0$ the set $L_\gamma(x)$ converges to the value of a Bayes rule at x. For example, with quadratic loss the Bayes rule is given by the posterior mean and a γ-lowest posterior loss region is the smallest sphere centered at the mean containing at least γ of the posterior probability.

Typically, in the continuous context, Bayes rules will not be invariant under reparameterizations. Robert (1996) recommended using the intrinsic loss function based on a measure of distance between sampling distributions as Bayes rules with respect to such losses are invariant. Bernardo (2005) recommended using the intrinsic loss function based on the Kullback-Leibler divergence $KL(f_\theta, f_{\theta'})$ between f_θ and $f_{\theta'}$. When $\psi = \theta$ the intrinsic loss function is given by $L(\theta, \theta') = \min(KL(f_\theta, f_{\theta'}), KL(f_{\theta'}, f_\theta))$. For a general marginal parameter ψ the intrinsic loss function is defined by $L(\theta, \psi) = \inf_{\theta' \in \Psi^{-1}\{\psi\}} L(\theta, \theta')$.

It can be shown, for example see Bernardo and Smith (2000) and Section 4, that hpd inferences arise as the limits of Bayes rules via a sequence of loss functions given by

$$L_\lambda(\theta, \psi) = I(\Psi(\theta) \notin B_\lambda(\psi)) \tag{5}$$

where $\lambda > 0$ and $B_\lambda(\Psi(\theta))$ is the ball of radius λ centered at ψ. As previously noted these inferences are not invariant under reparameterizations. It is our purpose here to show that relative surprise inferences also arise via a sequence of loss functions similar to (5) but based on the prior. So the loss functions are also in a sense intrinsic but based on the prior and not the sampling model, as with the intrinsic loss function.

In Section 2 we develop the prior-based loss function and show that ψ_{LRSE} is a Bayes rule when Ψ is finite. In Sections 3 and 4 we extend this result to show that ψ_{LRSE} is generally a limit of Bayes rules. In Section 5 we discuss

prediction problems and in Section 6 show that relative surprise regions are limits of γ-lowest posterior loss credible regions.

It is easy to see that the class of relative surprise credible regions $\{C_\gamma(x) : \gamma \in [0,1]\}$ for ψ is independent of the marginal prior π_Ψ. We note, however, that when we specify a $\gamma \in [0,1]$, the set $C_\gamma(x)$ does depend on π_Ψ through $c_\gamma(x)$. So the form of relative surprise inferences about ψ is completely robust to the choice of π_Ψ but the quantification of the uncertainty in the inferences is not. For example, when $\psi = \Psi(\theta) = \theta$, then $\theta_{\rm LRSE}(x)$ is the MLE while, in general, $\psi_{\rm LRSE}(x)$ is the maximizer of the integrated likelihood where we have integrated out nuisance parameters via the conditional prior given ψ. Similarly, relative surprise regions are likelihood regions in the case of the full parameter, and integrated likelihood regions generally. As such, the results derived in this paper establish that likelihood inferences are essentially Bayesian in character. We note, however, that a relative belief ratio $\pi_\Psi(\psi_0 \mid x)/\pi_\Psi(\psi_0)$, while proportional to an integrated likelihood, has an interpretation as a change in belief and cannot be multiplied by an arbitrary positive constant, as with a likelihood, without losing this interpretation.

In Le Cam (1953) it is shown that the MLE is asymptotically Bayes but this is for a fixed loss function, with increasing amounts of data and a sequence of priors. In this paper the amount of data and the prior are fixed but we may require a sequence of loss functions, to show that the MLE is a limit of Bayes rules. Berger, Liseo and Wolpert (1999) discuss maximum integrated likelihood estimates where default or noninformative priors are used to integrate out nuisance parameters and show good properties for this approach. Aitkin (2010) develops an approach to assessing hypotheses using the posterior distribution of likelihood ratios that is based on earlier work by Dempster (1973). As that approach does not use integrated likelihoods and, as of this time, doesn't have a decision-theoretic formulation, it is quite different than what we discuss here.

2 Estimation from Prior-based Loss Functions: The Finite Case

The following theorem presents the basic definition of the loss function when Ψ is finite and establishes an important optimality result. For more general situations we will need to modify this loss function slightly.

Theorem 1. Suppose that $\pi_\Psi(\psi) > 0$ for every $\psi \in \Psi$ and that Ψ is finite with ν_Ψ equal to counting measure. Then for the loss function

$$L(\theta, \psi) = \frac{I(\Psi(\theta) \neq \psi)}{\pi_\Psi(\Psi(\theta))} \tag{6}$$

a Bayes rule is given by $\psi_{\rm LRSE}$.
Proof: We have that

$$r(\delta \mid x) = \int_{\Theta} \frac{I(\Psi(\theta) \neq \delta(x))}{\pi_{\Psi}(\Psi(\theta))} \pi(\theta \mid x)\, \nu(d\theta) = \int_{\Psi} \frac{I(\psi \neq \delta(x))}{\pi_{\Psi}(\psi)} \pi_{\Psi}(\psi \mid x)\, \nu_{\Psi}(d\psi)$$
$$= \int_{\Psi} \frac{\pi_{\Psi}(\psi \mid x)}{\pi_{\Psi}(\psi)}\, \nu_{\Psi}(d\psi) - \frac{\pi_{\Psi}(\delta(x) \mid x)}{\pi_{\Psi}(\delta(x))}. \tag{7}$$

Since Ψ is finite, the first term in (7) is finite and a Bayes rule at x is given by the value $\delta(x)$ that maximizes the second term. Therefore, $\psi_{\mathrm{LRSE}}(x)$ is a Bayes rule.

From (7) the prior risk of δ is

$$r(\delta) = \#(\Psi) - E_M(\pi_{\Psi}(\delta(x) \mid x)/\pi_{\Psi}(\delta(x))) = \sum_{\psi} M_{\psi}(\delta(x) \neq \psi) \tag{8}$$

where E_M denotes expectation with respect to the prior predictive and M_{ψ} is the probability measure on $\mathcal{X}$ obtained by averaging P_{θ} using the conditional prior given that $\Psi(\theta) = \psi$, namely, $M_{\psi}(A) = \int_{\Psi^{-1}\{\psi\}} P_{\theta}(A)\, \Pi(d\theta \mid \Psi(\theta) = \psi)$. Therefore, finding a Bayes rule with respect to (6) is equivalent to finding δ that maximizes $E_M(\pi_{\Psi}(\delta(x) \mid x)/\pi_{\Psi}(\delta(x)))$. So a Bayes rule maximizes the prior expected relative belief ratio evaluated at the estimate and it is clear that the LRSE is a Bayes rule as it maximizes the relative belief ratio for each x.

If instead we take the loss function to be $I(\Psi(\theta) \neq \psi)$, then virtually the same proof establishes that ψ_{MAP} is a Bayes rule. The prior risk for this loss function and estimator δ can be written as

$$\sum_{\psi} M_{\psi}(\delta(x) \neq \psi) \pi_{\Psi}(\psi) \tag{9}$$

which is the prior probability of making an error. Both $I(\Psi(\theta) \neq \psi)$ and (6) are two-valued loss functions but, when we make an incorrect decision, the loss is constant in $\Psi(\theta)$ for $I(\Psi(\theta) \neq \psi)$ while it equals the reciprocal of the prior probability of $\Psi(\theta)$ for (6). So (6) penalizes an incorrect decision much more severely when the true value of $\Psi(\theta)$ is in the tails of the prior. This makes sense as we would want to override the effect of the prior when the prior is not placing appreciable mass at the true value. Note that $\psi_{\mathrm{MAP}} = \psi_{\mathrm{LRSE}}$ when Π_{Ψ} is uniform.

As we have already noted $\pi_{\Psi}(\psi \mid x)/\pi_{\Psi}(\psi)$ is proportional to the integrated likelihood of ψ when we integrate the likelihood with respect to the conditional prior of θ given ψ. So, under the conditions of Theorem 1, we have shown that the maximum integrated likelihood estimator is a Bayes rule. Furthermore, the Bayes rule is the same for every choice of π_{Ψ} and only depends on the full prior through the conditional prior placed on the nuisance parameters. When $\psi = \theta$ then $\psi_{\mathrm{LRSE}}(x)$ is the MLE of θ and so the MLE of θ is a Bayes rule for every prior π.

We consider an application.

Example 1. *Classification*

For a classification problem we have k categories $\{\psi_1, \ldots, \psi_k\}$ prescribed by some function Ψ, where $\pi_\Psi(\psi_i) > 0$ for each i. Based on observed data x we want to classify the data as having come from one of the distributions in the classes specified by $\Psi^{-1}\{\psi_i\}$.

The standard Bayesian solution to this problem is to use $\psi_{\mathrm{MAP}}(x)$ as the classifier. From (9) we have that $\psi_{\mathrm{MAP}}(x)$ minimizes the prior probability of misclassification. Note that $M_\psi(\delta(x) \neq \psi)$ is the prior probability of a misclassification given that ψ is the correct class and (9) is the weighted average of these probabilities where the weights are given by the prior probabilities of the ψ. We see from (8) that $\psi_{\mathrm{LRSE}}(x)$ is instead minimizing the sum over ψ of the probabilities of misclassification given that ψ is the correct class. So the essence of the difference between these two approaches in this problem is that $\psi_{\mathrm{LRSE}}(x)$ treats the errors of misclassification equally while $\psi_{\mathrm{MAP}}(x)$ weights them by their prior probabilities of occurrence.

We note that (8) is an upper bound on (9). So if the Bayes risk for loss function (6) is small, the prior risk of $\psi_{\mathrm{LRSE}}(x)$, with respect to the loss function $I(\Psi(\theta) \neq \psi)$, is also small, i.e., when using $\psi_{\mathrm{LRSE}}(x)$ the overall prior probability of a misclassification will also be small.

In general, it seems appropriate to be concerned with minimizing each of the probabilities $M_\psi(\delta(x) \neq \psi)$ and not downweight those corresponding to ψ values that have small prior probability. As a specific simple example suppose $k = 2$ and $x \sim$ Binomial(ψ_1) or $x \sim$ Binomial(ψ_2) with $\pi(\psi_1) = 1 - \epsilon$ and $\pi(\psi_2) = \epsilon$. After observing x we want to classify the observation. For example, ψ_i could be the probability of a diagnostic test for a disease indicating that the disease is present. We suppose that ψ_1 is the probability of a positive diagnostic test for the nondiseased population while ψ_2 is this probability for the diseased population. Further suppose that ψ_1/ψ_2 is very small, indicating that the test is successful in identifying the disease while not yielding many false positives, and suppose ϵ is very small, indicating that the disease is very rare. We have that $\pi(\psi_1 \mid 1) = \psi_1(1-\epsilon)/(\psi_1(1-\epsilon)+\psi_2\epsilon)$ and $\pi(\psi_1 \mid 0) = (1-\psi_1)(1-\epsilon)/((1-\psi_1)(1-\epsilon) + (1-\psi_2)\epsilon)$. Therefore, $\psi_{\mathrm{MAP}}(1) = \psi_1$ if $\psi_1/\psi_2 > \epsilon/(1-\epsilon)$ and is ψ_2 otherwise, while $\psi_{\mathrm{MAP}}(0) = \psi_1$ if $(1-\psi_1)/(1-\psi_2) > \epsilon/(1-\epsilon)$ and is ψ_2 otherwise. Also $\psi_{\mathrm{LRSE}}(1) = \psi_1$ if $\psi_1 > \psi_2$ and is ψ_2 otherwise, while $\psi_{\mathrm{LRSE}}(0) = \psi_1$ if $(1-\psi_1) > (1-\psi_2)$ and is ψ_2 otherwise. So we see from this that ψ_{MAP} will always classify a person to the nondiseased population when ϵ is small enough, e.g., take $\psi_1 = 0.05, \psi_2 = 0.80$, and $\epsilon < 0.0566$. By contrast, in this situation, ψ_{LRSE} will always classify an individual with a positive test to the diseased population and to the nondiseased population for a negative test. Now M_{ψ_i} is the Binomial(ψ_i) distribution, so when $\psi_1 < \psi_2$ and ϵ is small enough

$$M_{\psi_1}(\psi_{\mathrm{MAP}} \neq \psi_1) + M_{\psi_2}(\psi_{\mathrm{MAP}} \neq \psi_2) = 0 + 1 = 1,$$
$$M_{\psi_1}(\psi_{\mathrm{LRSE}} \neq \psi_1) + M_{\psi_2}(\psi_{\mathrm{LRSE}} \neq \psi_2) = \psi_1 + (1-\psi_2) < 1.$$

This illustrates clearly the difference between these two procedures as ψ_{LRSE}

does vastly better than ψ_{MAP} on the diseased population when ψ_1 is small and ψ_2 is large as would be the case for a good diagnostic. Of course ψ_{MAP} minimizes the overall error rate but at the price of ignoring the most important class in this problem. Note that this example can be extended to the situation where we need to estimate the ψ_i based on samples from the respective populations but this will not materially affect the overall conclusions. Also see Example 3 where ϵ is considered unknown.

In a general estimation problem an estimator δ is unbiased with respect to a loss function L if $E_\theta(L(\theta', \delta(x))) \geq E_\theta(L(\theta, \delta(x)))$ for all $\theta', \theta \in \Theta$. This says that on average $\delta(x)$ is closer to the true value than any other value when we interpret $L(\theta, \delta(x))$ as a measure of distance between the estimate and what is being estimated. A reasonable definition of *Bayesian unbiasedness* for δ with respect to L is thus obtained by requiring that

$$\int_\Theta \int_\Theta E_\theta(L(\theta', \delta(x)))\, \Pi(d\theta)\, \Pi(d\theta') \geq \int_\Theta E_\theta(L(\theta, \delta(x)))\, \Pi(d\theta) = r(\delta).$$

Here we are thinking of θ' as a false value generated from the prior independently of the true value θ so θ' has no connection with the data. Therefore, δ is Bayesian unbiased if on average $\delta(x)$ is closer to the true value than a false value. In Section 3 we prove that ψ_{LRSE} is Bayesian unbiased with respect to a general class of loss functions that includes both (6) and $I(\Psi(\theta) \neq \psi)$.

3 Estimation from Prior-based Loss Functions: The Countably Infinite Case

The loss function (6) does not provide meaningful results when Ψ is infinite as (8) shows that $r(\delta)$ will be infinite. So we modify (6) via a parameter $\eta > 0$ and define the loss function

$$L_\eta(\theta, \psi) = \frac{I(\Psi(\theta) \neq \psi)}{\max(\eta, \pi_\Psi(\Psi(\theta)))} \tag{10}$$

and note that L_η is a bounded function of (θ, ψ). This loss function is like (6) but does not allow for arbitrarily large losses. Without loss of generality we can restrict η to a sequence of values converging to 0. We prove the following result in the Appendix.

Theorem 2. Suppose that $\pi_\Psi(\psi) > 0$ for every $\psi \in \Psi$, that Ψ is countable with ν_Ψ equal to counting measure and that $\psi_{\mathrm{LRSE}}(x)$ is the unique maximizer of $\pi_\Psi(\psi \,|\, x)/\pi_\Psi(\psi)$ for all x. For the loss function (10) and Bayes rule δ_η, then $\delta_\eta(x) \to \psi_{\mathrm{LRSE}}(x)$ as $\eta \to 0$, for every $x \in \mathcal{X}$.

The proof of Theorem also establishes the following result.

Corollary 3. For all sufficiently small η the value of the Bayes rule at x is given by $\psi_{\mathrm{LRSE}}(x)$.

If instead we take the loss function to be $I(\Psi(\theta) \neq \psi)$, then virtually the same proof as in Theorem 1 establishes that ψ_{MAP} is a Bayes rule.

We now investigate the unbiasedness of $\psi_{\mathrm{LRSE}}(x)$. For this we consider loss functions of the form

$$L(\theta, \psi) = I(\Psi(\theta) \neq \psi) h(\Psi(\theta)) \tag{11}$$

for some nonnegative function h which satisfies $\int_\Theta h(\Psi(\theta))\, \Pi(d\theta) < \infty$. This class of loss functions includes (6) when Ψ is finite, (10) and $I(\Psi(\theta) \neq \psi)$. We have the following result.

Theorem 4. If Ψ is countable, then $\psi_{\mathrm{LRSE}}(x)$ is Bayesian unbiased under the loss function (11).

Proof: The prior risk of δ is given by

$$\begin{aligned} r(\delta) &= \int_\Theta \int_{\mathcal{X}} L(\theta, \delta(x))\, P_\theta(dx)\, \Pi(d\theta) \\ &= \int_\Theta \int_{\mathcal{X}} [h(\Psi(\theta)) - I(\Psi(\theta) = \delta(x)) h(\Psi(\theta))]\, P_\theta(dx)\, \Pi(d\theta) \\ &= \int_\Theta h(\Psi(\theta))\, \Pi(d\theta) - \int_{\mathcal{X}} \int_\Theta I(\Psi(\theta) = \delta(x)) h(\Psi(\theta))\, \Pi(d\theta \,|\, x)\, M(dx) \\ &= \int_\Theta h(\Psi(\theta))\, \Pi(d\theta) - \int_{\mathcal{X}} h(\delta(x)) \pi_\Psi(\delta(x) \,|\, x)\, M(dx) \end{aligned}$$

and

$$\begin{aligned} &\int_\Theta \int_\Theta \int_{\mathcal{X}} L(\theta', \delta(x))\, P_\theta(dx)\, \Pi(d\theta)\, \Pi(d\theta') \\ &= \int_\Theta \int_\Theta \int_{\mathcal{X}} [h(\Psi(\theta')) - I(\Psi(\theta') = \delta(x)) h(\Psi(\theta'))]\, P_\theta(dx)\, \Pi(d\theta)\, \Pi(d\theta') \\ &= \int_\Theta h(\Psi(\theta))\, \Pi(d\theta) - \int_{\mathcal{X}} h(\delta(x)) \pi_\Psi(\delta(x))\, M(dx). \end{aligned}$$

Therefore, δ is Bayesian unbiased if and only if

$$\int_{\mathcal{X}} h(\delta(x)) [\pi_\Psi(\delta(x) \,|\, x) - \pi_\Psi(\delta(x))]\, M(dx) \geq 0. \tag{12}$$

It is a consequence of results proved in Evans and Shakhatreh (2008) that it is always true that $\pi_\Psi(\psi_{\mathrm{LRSE}}(x) \,|\, x) / \pi_\Psi(\psi_{\mathrm{LRSE}}(x)) \geq 1$ and this establishes the result. This can also be seen by noting that $\pi_\Psi(\cdot \,|\, x)/\pi_\Psi(\cdot)$ is the density of $\Pi_\Psi(\cdot \,|\, x)$ with respect to Π_Ψ and so we must have that the maximum of this density is greater than or equal to 1.

The proof gives a sufficient condition for Bayesian unbiasedness with respect to the loss (11).

Corollary 5. δ is Bayesian unbiased if $\pi_\Psi(\delta(x) \,|\, x) \geq \pi_\Psi(\delta(x))$ for all x.

At this point we have neither a proof of the Bayesian unbiasedness of ψ_{MAP} with respect to $I(\Psi(\theta) \neq \psi)$, nor a counterexample although we suspect that

it is not. We do know, however, that $\psi_{\rm MAP}$ is Bayesian unbiased with respect to $I(\Psi(\theta) \neq \psi)$ whenever Π_Ψ is uniform because in that case $\psi_{\rm MAP} = \psi_{\rm LRSE}$. It is also clear from (11) that $\psi_{\rm LRSE}$ possesses a very strong property as the integrand is always nonnegative when $\delta = \psi_{\rm LRSE}$. In light of this we refer to an estimator possessing this property as being *uniformly (in x) Bayesian unbiased*.

4 Estimation from Prior-based Loss Functions: The Continuous Case

When ψ has a continuous prior distribution the argument in Theorem 2 does not work as $\Pi_\Psi(\{\delta(x)\} \,|\, x) = 0$. There are several possible ways to proceed here but we consider a discretization of the problem that uses Theorem 2. For this we will assume that the spaces involved are locally Euclidean, mappings are sufficiently smooth and take the support measures to be the analogs of Euclidean volume on the respective spaces. Further details on the mathematical requirements underlying these assumptions can be found in Tjur (1974) where spaces are taken to be Riemann manifolds. While the argument we provide applies quite generally, we simplify this here by taking all spaces to be open subsets of Euclidean spaces and the support measures to be Euclidean volume on these sets.

For each $\lambda > 0$ we discretize the set Ψ via a countable partition $\{B_\lambda(\psi) : \psi \in \Psi\}$ where $\psi \in B_\lambda(\psi), \Pi_\Psi(B_\lambda(\psi)) > 0$, $\sup_{\psi\in\Psi} {\rm diam}(B_\lambda(\psi)) \to 0$ as $\lambda \to 0$. For example, the $B_\lambda(\psi)$ could be equal volume rectangles in R^k. Further, we assume that $\Pi_\Psi(B_\lambda(\psi))/\nu_\Psi(B_\lambda(\psi)) \to \pi_\Psi(\psi)$ as $\lambda \to 0$ for every ψ. This will hold whenever π_Ψ is continuous everywhere and $B_\lambda(\psi)$ converges nicely to $\{\psi\}$ as $\lambda \to 0$ (see Rudin (1974), Chapter 8 for the definition of 'converges nicely'). Let $\psi_\lambda(\psi) \in B_\lambda(\psi)$ be such that $\psi_\lambda(\psi') = \psi_\lambda(\psi)$ whenever $\psi' \in B_\lambda(\psi)$ and $\Psi_\lambda = \{\psi_\lambda(\psi) : \psi_\lambda(\psi) \in B_\lambda(\psi)\}$ be the discretized version of Ψ. Note that one point is chosen in each $B_\lambda(\psi)$. We will call this a *regular discretization* of Ψ. The discretized prior on Ψ_λ is $\pi_{\Psi,\lambda}(\psi_\lambda(\psi)) = \Pi_\Psi(B_\lambda(\psi))$ and the discretized posterior is $\pi_{\Psi,\lambda}(\psi_\lambda(\psi) \,|\, x) = \Pi_\Psi(B_\lambda(\psi) \,|\, x)$.

We define the loss function for the discretized problem just as for Theorem 2, by

$$L_{\lambda,\eta}(\theta, \psi_\lambda(\psi)) = \frac{I(\psi_\lambda(\Psi(\theta)) \neq \psi_\lambda(\psi))}{\max(\eta, \pi_{\Psi,\lambda}(\psi_\lambda(\Psi(\theta))))} \tag{13}$$

and denote a Bayes rule for this problem by $\delta_{\lambda,\eta}(x)$. In this case we not only need that $\psi_{\rm LRSE}(x)$ is the unique maximizer of $\pi_\Psi(\psi \,|\, x)/\pi_\Psi(\psi)$, but we cannot allow $\pi_\Psi(\psi \,|\, x)/\pi_\Psi(\psi)$ to come arbitrarily close to its maximum outside a neighborhood of $\psi_{\rm LRSE}(x)$. It is clear that when this does not hold then we are in a pathological situation that will not apply in a typical application. The following result is proved in the Appendix.

Theorem 6. Suppose that π_Ψ is positive and continuous and we have a regular discretization of Ψ. Further suppose that $\psi_{\rm LRSE}(x)$ is the unique maximizer of $\pi_\Psi(\psi \,|\, x)/\pi_\Psi(\psi)$ and for any $\epsilon > 0$

$$\sup_{\{\psi:||\psi-\psi_{\rm LRSE}(x)||\geq\epsilon\}} \frac{\pi_\Psi(\psi\,|\,x)}{\pi_\Psi(\psi)} < \frac{\pi_\Psi(\psi_{\rm LRSE}(x)\,|\,x)}{\pi_\Psi(\psi_{\rm LRSE}(x))}.$$

Then, there exists $\eta(\lambda) > 0$ such that a Bayes rule $\delta_{\lambda,\eta(\lambda)}(x)$ converges to $\psi_{\rm LRSE}(x)$ as $\lambda \to 0$ for all x.

Theorem 6 says that $\psi_{\rm LRSE}$ is a limit of Bayes rules. So when $\Psi(\theta) = \theta$ we have the result that the MLE is a limit of Bayes rules and more generally the maximum integrated likelihood estimator is a limit of Bayes rules.

Now let $\hat{\psi}_\lambda(x)$ be the LRSE of the discretized problem, i.e., $\hat{\psi}_\lambda(x)$ maximizes $\Pi_\Psi(B_\lambda(\psi)\,|\,x)/\Pi_\Psi(B_\lambda(\psi))$ as a function of $\psi \in \Psi_\lambda$. The following result is proved in the Appendix.

Corollary 7. $\hat{\psi}_\lambda$ converges to $\psi_{\rm LRSE}$ as $\lambda \to 0$.

Note that by Theorem 4, $\hat{\psi}_\lambda$ is uniformly Bayesian unbiased for the discretized problem. Therefore, $\psi_{\rm LRSE}$ is the limit of uniformly Bayesian unbiased estimators.

By similar arguments we can establish an analog of Theorem 6 for $\psi_{\rm MAP}$ using the loss function given by (5). Actually in this case a simpler development can be followed in certain situations. For this note that the posterior risk of δ is given by $1 - \Pi_\Psi(B_\lambda(\delta(x))\,|\,x) = 1 - \pi_\Psi(\delta'(x)\,|\,x)\nu_\Psi(B_\lambda(\delta(x)))$ for some $\delta'(x) \in B_\lambda(\delta(x))$. Now suppose we take $B_\lambda(\psi)$ to be a sphere of radius λ centered at ψ. Suppose further that for each $\epsilon > 0$ there exists a $\lambda(\epsilon) > 0$ such that when $||\psi - \psi_{\rm MAP}(x)|| > \lambda(\epsilon)$ then $\pi_\Psi(\psi\,|\,x) < \inf_{\psi' \in B_{\lambda(\epsilon)}(\psi_{\rm MAP}(x))} \pi_\Psi(\psi'\,|\,x)$. Since $\nu_\Psi(B_\lambda(\psi))$ is constant we have that a Bayes rule δ_λ must then satisfy $||\delta_\lambda(x) - \psi_{\rm MAP}(x)|| < \epsilon$. So we have proved that $\psi_{\rm MAP}$ is a limit of Bayes rules. By contrast, for the loss function $I(\Psi(\theta) \notin B_\lambda(\psi))/\Pi_\Psi(B_\lambda(\Psi(\theta)))$ the posterior risk of δ is given by $\int_\Psi \{\Pi_\Psi(B_\lambda(\psi))\}^{-1}\Pi_\Psi(d\psi\,|\,x) - \int_{B_\lambda(\delta(x))} \{\Pi_\Psi(B_\lambda(\psi))\}^{-1}\Pi_\Psi(d\psi\,|\,x)$. The simpler approach is not available in this case because the first term is unbounded.

We consider now an important example.

Example 2. *Regression (estimation)*

Suppose that we have $y = X\beta + e$ where $y \in R^n, X \in R^{n\times k}$ is fixed, $\beta \in R^{n\times k}$, and $e \sim N_n(0, \sigma^2 I)$. We will assume that σ^2 is known to simplify the discussion. Let π be a prior density for β. Then having observed (X, y), $\beta_{\rm LRSE}(x) = b = (X'X)^{-1}X'y$ which is the MLE of β. It is interesting to contrast this result with what might be considered more standard Bayesian estimates such as the posterior mode or posterior mean. For example, suppose that $\beta \sim N_k(0, \tau^2 I)$. Then the posterior distribution of β is $N_k(\mu_{\rm post}(\beta), \Sigma_{\rm post}(\beta))$ where

$$\mu_{\rm post}(\beta) = \Sigma_{\rm post}(\beta)\sigma^{-2}X'Xb,\ \Sigma_{\rm post}(\beta) = (\tau^{-2}I + \sigma^{-2}X'X)^{-1}$$

and the posterior mean and modal estimates of β are both equal to $\mu_{\rm post}(\beta)$. Writing the spectral decomposition of $X'X$ as $X'X = Q\Lambda Q'$ we have that

$$||\mu_{\rm post}(\beta)|| = ||(I + (\sigma^2/\tau^2)\Lambda^{-1})^{-1}Q'b||.$$

Since $||b|| = ||Q'b||$ and $1/(1+\sigma^2/(\tau^2\lambda_i)) < 1$ for each i, we see that $\mu_{\text{post}}(\beta)$ moves the MLE towards the prior mean 0. This is often cited as a positive attribute of these estimates but consider the situation where the true value of β lies in the tails of the prior. In that case it is certainly wrong to move β towards the prior mean. When τ^2 is chosen very large, so we avoid the possibility that the true value of β lies in the tails of the prior, then the MLE and the posterior mean are virtually the same. It makes sense to choose $\tau^2 > \sigma^2$ as this says we have less prior information about a β_i than the amount we learn about β from a single observation. So it is not clear that shrinking the MLE is necessarily a good thing particularly as this requires giving up invariance.

Suppose now we want to estimate $\psi = w'\beta$ for some setting w of the predictors. The prior distribution of ψ is $N(0, \sigma^2_{\text{prior}}(\psi)) = N(0, \tau^2 w'w)$ and the posterior distribution is $N(\mu_{\text{post}}(\psi), \sigma^2_{\text{post}}(\psi)) = N(w'\mu_{\text{post}}(\beta), x'\Sigma_{\text{post}}(\beta)x)$. Note that $\sigma^2_{\text{prior}}(\psi) - \sigma^2_{\text{post}}(\psi) = w'(\tau^2 I - \Sigma_{\text{post}}(\beta))w = \tau^2 w'Q'(I - (I + (\tau^2/\sigma^2)\Lambda)^{-1})Qw > 0$ and so maximizing the ratio of the posterior to prior densities leads to

$$\psi_{\text{LRSE}}(y) = (1 - \sigma^2_{\text{post}}(\psi)/\sigma^2_{\text{prior}}(\psi))^{-1}\mu_{\text{post}}(\psi). \tag{14}$$

Since $\sigma^2_{\text{prior}}(\psi) > \sigma^2_{\text{post}}(\psi)$ we have $|\psi_{\text{LRSE}}(y)| > |\mu_{\text{post}}(\psi)|$ and $\mu_{\text{post}}(\psi) = \psi_{\text{MAP}}(y)$. Note that when $\sigma^2_{\text{post}}(\psi)$ is much smaller than $\sigma^2_{\text{prior}}(\psi)$, in other words the posterior is densely concentrated about $\mu_{\text{post}}(\psi)$, then $w_{\text{LRSE}}(y)$ and $w_{\text{MAP}}(y)$ are very similar. In general $\psi_{\text{LRSE}}(y)$ is not equal to $w'b$, the plug-in MLE of ψ, although $\psi_{\text{LRSE}}(y) \to w'b$ as $\tau^2 \to \infty$.

5 Prediction from Prior-based Loss Functions

Suppose after observing x we want to predict a future value $y \in \mathcal{Y}$ where y has model given by $g_{\eta(\theta)}(y \,|\, x)$ with respect to support measure $\mu_{\mathcal{Y}}$ on $\mathcal{Y}$. We allow for the possibility here that the distribution of y depends on x and also that θ may not index these distributions. Then we have that the joint density of (θ, x, y) is given by $\pi(\theta)f_\theta(x)g_{\eta(\theta)}(y \,|\, x)$ and after observing x the conditional density of y is given by the posterior predictive density $q(y \,|\, x) = \int_\Theta \pi(\theta \,|\, x) g_{\eta(\theta)}(y \,|\, x)\, \nu(d\theta)$ while the prior predictive density of y is given by $q(y) = \int_\Theta \int_{\mathcal{X}} \pi(\theta) f_\theta(x) g_{\eta(\theta)}(y \,|\, x)\, \mu(dx)\, \nu(d\theta)$. Therefore, the relative belief in a future value y is given by $q(y \,|\, x)/q(y)$ and we denote the maximizer of this by $y_{\text{LRSE}}(x)$.

Again the LRSE arises from loss function considerations. For example, when $\mathcal{Y}$ is finite we consider the loss function

$$L(y, y') = \frac{I(y \neq y')}{q(y)},$$

where we think of y as some true value of y that is concealed from us by the future, or some other mechanism, and which we want to predict. Then the

posterior risk of a predictor $\delta : \mathcal{X} \to \mathcal{Y}$ is given by

$$r(\delta \,|\, x) = \int_{\mathcal{Y}} \frac{q(y \,|\, x)}{q(y)} \, \mu_{\mathcal{Y}}(dy) - \frac{q(\delta(x) \,|\, x)}{q(\delta(x))}$$

and we see that y_{LRSE} is a Bayes rule. Also, the prior risk of predictor δ is given by $r(\delta) = \sum_y M_y(\delta(x) \neq y)$ where M_y is the conditional prior predictive of x given y and so $r(\delta)$ is the sum of the conditional prediction errors given y. We can also develop results similar to Theorems 2 and 6 for the situation where $\mathcal{Y}$ is not finite to show that y_{LRSE} is a limit of Bayes rules.

We consider some examples.

Example 3. *Classification (prediction)*

Consider now a situation where (x, c) is such that $x \,|\, c \sim f_c$ with $c \sim$ Bernoulli(ϵ) where f_0 and f_1 are known (or accurately estimated based on large samples) but ϵ is unknown with prior π. This is a generalization of Example 1 where ϵ was assumed to be known. Then based on a sample $(x_1, c_1), \ldots, (x_n, c_n)$ from the joint distribution we want to predict the value c_{n+1} for a newly observed x_{n+1}. Therefore, $q(c) = \int_0^1 (1-\epsilon)^{1-c} \epsilon^c \pi(\epsilon) \, d\epsilon$ and, if $\epsilon \sim \text{Beta}(\alpha, \beta)$, the prior predictive of c_{n+1} is Bernoulli$(\alpha/(\alpha+\beta))$. For c_{n+1} the posterior predictive density is $q(c \,|\, (x_1, c_1), \ldots, (x_n, c_n), x_{n+1}) \propto (f_0(x_{n+1}))^{1-c} (f_1(x_{n+1}))^c \int_0^1 \epsilon^{n\bar{c}+c} (1 - \epsilon)^{n(1-\bar{c})+(1-c)} \pi(\epsilon) \, d\epsilon$ with $\bar{c} = n^{-1} \sum_{i=1}^n$
c_i. With a Beta(α, β) prior for ϵ, we have that $q(c \,|\, (x_1, c_1), \ldots, (x_1, c_1), x_{n+1})$
$\propto f_c(x_{n+1}) \Gamma(\alpha + n\bar{c} + c) \, \Gamma(\beta + n(1-\bar{c}) + 1 - c)$. From this we see immediately that

$$c_{\text{MAP}} = \begin{cases} 1 & \text{if } \frac{f_1(x_{n+1})}{f_0(x_{n+1})} \frac{(\alpha+n\bar{c})}{(\beta+n(1-\bar{c}))} \geq 1 \\ 0 & \text{otherwise,} \end{cases}$$

$$c_{\text{LRSE}} = \begin{cases} 1 & \text{if } \frac{f_1(x_{n+1})}{f_0(x_{n+1})} \frac{\beta(\alpha+n\bar{c})}{\alpha(\beta+n(1-\bar{c}))} \geq 1 \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

Note that c_{MAP} and c_{LRSE} are identical whenever $\alpha = \beta$.

We can see from these formulas that a substantial difference will arise between c_{MAP} and c_{LRSE} when one of α or β is much bigger than the other. As in Example 1 these correspond to situations where we believe that ϵ or $1-\epsilon$ is very small. Suppose we take $\alpha = 1$ and let β be relatively large, as this corresponds to knowing *a priori* that ϵ is very small. Then (15) implies that $c_{\text{MAP}} \leq c_{\text{LRSE}}$ and so $c_{\text{LRSE}} = 1$ whenever $c_{\text{MAP}} = 1$. A similar conclusion arises when we take $\beta = 1$ and $\alpha < 1$.

To see what kind of improvement is possible we consider a simulation. Here we take f_0 to be a $N(0, 1)$ density, f_1 to be a $N(\mu, 1)$ density, let $n = 10$ and the prior on ϵ be Beta$(1, \beta)$. Table 1 presents the Bayes risks for c_{MAP} and c_{LRSE} for various choices of β when $\mu = 1$. When $\beta = 1$ they are equivalent but we see that as β rises the performance of c_{MAP} deteriorates while c_{LRSE} improves. Large values of β correspond to having information that ϵ is small. When $\beta = 14$ about 0.50 of the prior probability is to the left of 0.05, with $\beta = 32$ about 0.80

β	$M_0(c_{\text{MAP}} \neq 0) + M_1(c_{\text{MAP}} \neq 1)$	$M_0(c_{\text{LRSE}} \neq 0) + M_1(c_{\text{LRSE}} \neq 1)$
1	$0.386 + 0.390 = 0.776$	$0.386 + 0.390 = 0.776$
14	$0.002 + 0.975 = 0.977$	$0.285 + 0.380 = 0.665$
32	$0.000 + 0.997 = 0.997$	$0.292 + 0.349 = 0.641$
100	$0.000 + 1.000 = 1.000$	$0.300 + 0.324 = 0.624$

Table 1: Conditional prior probabilities of misclassification for MAP and LRSE for various values of β in Example 3 when $\alpha = 1$, $\mu = 1$, and n=10.

of the prior probability is to the left of 0.05, and with $\beta = 100$ about 0.99 of the prior probability is to the left of 0.05. We see that the misclassification rates for the small group ($c = 1$) stay about the same for c_{LRSE} as β increases while they deteriorate markedly for c_{MAP} as the MAP procedure basically ignores the small group.

We also investigated other choices for n and μ. There is very little change as n increases. When μ moves towards 0 the error rates go up and go down as μ moves away from 0, as one would expect. Of course, c_{LRSE} always dominates c_{MAP}.

Example 4. *Regression (prediction)*

Consider the situation of Example 2 and suppose we want to predict a response z at the predictor value $w \in R^k$. When $\beta \sim N_k(0, \tau^2 I)$ the prior distribution of z is $z \sim N(0, \sigma^2 + \tau^2 w'w) = N(0, \sigma^2_{\text{prior}}(z))$ and the posterior distribution is $N(\mu_{\text{post}}(z), \sigma^2_{\text{post}}(z))$ where

$$\mu_{\text{post}}(z) = w'\mu_{\text{post}}(\beta),\ \sigma^2_{\text{post}}(z) = \sigma^2 + w'\Sigma_{\text{post}}(\beta)w.$$

To obtain $z_{\text{LRSE}}(y)$ we need to maximize the ratio of the posterior to the prior density of z and an easy calculation shows that this leads to

$$z_{\text{LRSE}}(y) = (1 - \sigma^2_{\text{post}}(z)/\sigma^2_{\text{prior}}(z))^{-1}\mu_{\text{post}}(z). \tag{16}$$

Note that $\sigma^2_{\text{prior}}(z) - \sigma^2_{\text{post}}(z) = \sigma^2_{\text{prior}}(w'\beta) - \sigma^2_{\text{post}}(w'\beta) > 0$ and so $|z_{\text{LRSE}}(y)| > |\mu_{\text{post}}(z)|$ and the LRSE is further from the prior mean than $z_{\text{MAP}}(y) = \mu_{\text{post}}(z)$. Also, we see that, when $\sigma^2_{\text{post}}(z)$ is small then $z_{\text{LRSE}}(y)$ and $z_{\text{MAP}}(y)$ are very similar. Finally, comparing (14) and (16) we have that

$$z_{\text{LRSE}}(y) = (\sigma^2_{\text{prior}}(z)/\sigma^2_{\text{post}}(\psi))w'\psi_{\text{LRSE}}(y) = (1 + \sigma^2/\tau^2)\psi_{\text{LRSE}}(y)$$

and so the LRSE predictor at x is more dispersed than the LRSE estimator of the mean at w and this makes good sense as we have to take into account the additional variation due to prediction. By contrast $w_{\text{MAP}}(y) = \psi_{\text{MAP}}(y)$.

6 Regions from Prior-based Loss Functions

We now consider the lowest posterior loss γ-credible regions that arise from the prior-based loss functions we have considered. Let $C_\gamma(x)$ denote a γ-relative

surprise region for ψ. Consider first the case where Ψ is finite. We have the following result.

Theorem 8. Suppose that $\pi_\Psi(\psi) > 0$ for every $\psi \in \Psi$ and that Ψ is finite with ν_Ψ equal to counting measure. Then for the loss function given by (6), $C_\gamma(x)$ is a γ-lowest posterior loss credible region.
Proof: From (4) and (7) the γ-lowest posterior loss credible region is

$$L_\gamma(x) = \left\{\psi : \frac{\pi_\Psi(\psi \,|\, x)}{\pi_\Psi(\psi)} \geq \int_\Psi \frac{\pi_\Psi(\zeta \,|\, x)}{\pi_\Psi(\zeta)}\, \nu_\Psi(d\zeta) - l_\gamma(x)\right\}$$

and $l_\gamma(x) = \inf\{k : \Pi_\Psi(\{\psi : r(\psi \,|\, x) \leq k\} \geq \gamma\}$. As $\int_\Psi (\pi_\Psi(z \,|\, x)/\pi_\Psi(z))\nu_\Psi(dz)$ is independent of ψ it is clearly equivalent to define this region via $C_\gamma(x) = \{\psi : \pi_\Psi(\psi \,|\, x)/\pi_\Psi(\psi) \geq c_\gamma(x)\}$, namely, $L_\gamma(x) = C_\gamma(x)$.

Now consider the case where Ψ is countable and we use loss function (10). Following the proof of Theorem 8 we see that a γ-lowest posterior loss region takes the form

$$L_{\eta,\gamma}(x) = \{\psi : \pi_\Psi(\psi \,|\, x)/\max(\eta, \pi_\Psi(\psi)) \geq l_{\eta,\gamma}(x)\}$$

where $l_{\eta,\gamma}(x) = \sup\{k : \Pi_\Psi(\{\psi : \pi_\Psi(\psi \,|\, x)/\max(\eta, \pi_\Psi(\psi)) \geq k\} \geq \gamma\}$. We prove the following result in the Appendix.

Theorem 9. Suppose that $\pi_\Psi(\psi) > 0$ for every $\psi \in \Psi$, that Ψ is countable with ν_Ψ equal to counting measure. For the loss function (10), we have that $C_\gamma(x) \subset \liminf_{\eta\to 0} L_{\eta,\gamma}(x)$ whenever γ is such that $\Pi_\Psi(C_\gamma(x) \,|\, x) = \gamma$ and $\limsup_{\eta\to 0} L_{\eta,\gamma}(x) \subset C_{\gamma'}(x)$ whenever $\gamma' > \gamma$ and $\Pi_\Psi(C_{\gamma'}(x) \,|\, x) = \gamma'$.

While Theorem 9 does not establish the exact convergence $\lim_{\eta\to 0} L_{\eta,\gamma}(x) = C_\gamma(x)$ we suspect, however, that this does hold under quite general circumstances due to the discreteness. Theorem 9 does show that limit points of the class of sets $L_{\eta,\gamma}(x)$ always contain $C_\gamma(x)$ and their posterior probability content differs from γ by at most $\gamma' - \gamma$ where $\gamma' > \gamma$ is the next largest value for which we have exact content.

We now consider the continuous case and suppose we have a regular discretization. For $S^* \subset \Psi_\lambda = \{\psi_\lambda(\psi) : \psi_\lambda(\psi) \in B_\lambda(\psi)\}$, namely, S^* is a subset of a discretized version of Ψ, we define the *undiscretized* version of S^* to be $S = \cup_{\psi\in S^*} B_\lambda(\psi)$. Now let $C^*_{\lambda,\gamma}(x)$ be the γ-relative surprise region for the discretized problem and let $C_{\lambda,\gamma}(x)$ be its undiscretized version. Note that in a continuous context we will consider two sets as equal if they differ only by a set of measure 0 with respect to Π_Ψ. In the Appendix we prove the following which says that a γ-relative surprise region for the discretized problem (after undiscretizing) converges to the γ-relative surprise region for the original problem.

Theorem 10. Suppose that π_Ψ is positive and continuous, we have a regular discretization of Ψ and $\pi_\Psi(\psi \,|\, x)/\pi_\Psi(\psi)$ has a continuous posterior distribution. Then $\lim_{\lambda\to 0} C_{\lambda,\gamma}(x) = C_\gamma(x)$.

While Theorem 10 has interest in its own right, we can use it to prove that relative surprise regions are limits of lowest posterior loss regions.

Let $L^*_{\eta,\lambda,\gamma}(x)$ be the γ-lowest posterior loss region obtained for the discretized problem using loss function (13) and let $L_{\eta,\lambda,\gamma}(x)$ be the undiscretized version. We prove the following result in the Appendix.

Theorem 11. Suppose that π_Ψ is positive and continuous, we have a regular discretization of Ψ and $\pi_\Psi(\psi \mid x)/\pi_\Psi(\psi)$ has a continuous posterior distribution. Then $C_\gamma(x) = \lim_{\lambda\to 0} \liminf_{\eta\to 0} L_{\eta,\lambda,\gamma}(x) = \lim_{\lambda\to 0} \limsup_{\eta\to 0} L_{\eta,\lambda,\gamma}(x)$.

In Evans, Guttman, and Swartz (2006) and Evans and Shakhatreh (2008) additional properties of relative surprise regions are developed. For example, it is proved that a γ-relative surprise region $C_\gamma(x)$ for ψ satisfying $\Pi_\Psi(C_\gamma(x) \mid x) = \gamma$ minimizes $\Pi_\Psi(B)$ among all (measurable) subsets of Ψ satisfying $\Pi_\Psi(B \mid x) \geq \gamma$. So a γ-relative surprise region is smallest among all γ-credible regions for ψ where size is measured using the prior measure. This property has several consequences. For example, the prior probability that a region $B(x) \subset \Psi$ contains a false value from the prior is given by $\int_\Theta \int_\Psi P_\theta(\psi \in B(x))\, \Pi_\Psi(d\psi)\, \Pi(d\theta)$ where a false value is a value of $\psi \sim \Pi_\Psi$ generated independently of $(\theta, x) \sim \Pi_\Psi \times P_\theta$. It can be proved that a γ-relative surprise region minimizes this probability among all γ-credible regions for ψ and is always unbiased in the sense that the probability of covering a false value is bounded above by γ. Furthermore, a γ-relative surprise region maximizes the relative belief ratio $\Pi_\Psi(B \mid x)/\Pi_\Psi(B)$ and the Bayes factor $\Pi_\Psi(B \mid x)\Pi_\Psi(B^c)/\Pi_\Psi(B^c \mid x)\Pi_\Psi(B)$ among all regions $B \subset \Psi$ with $\Pi_\Psi(B) = \Pi_\Psi(C_\gamma(x) \mid x)$.

While the results in this section have been concerned with obtaining credible regions for parameters, similar results can be proved for the construction of prediction regions.

7 Conclusions

Relative surprise inferences are closely related to likelihood inferences. This together with their invariance and optimality properties make these prime candidates as appropriate inferences in Bayesian contexts. This paper has shown that relative surprise inferences arise naturally in a decision-theoretic formulation using loss functions based on the prior. As of yet these inferences are not typically used while MAP-based inferences, which seem to possess few strong properties, are commonly recommended. Based on the properties we have discussed in this paper we conclude that improvements in inferences can be accomplished by adopting relative surprise inferences. While we have required proper priors in this paper, limiting relative surprise inferences, as priors become increasingly diffuse, can also be obtained and have been discussed in the references.

Relative surprise estimation of the parameter ψ is based on the relative belief ratio $\pi_\Psi(\psi \mid x)/\pi_\Psi(\psi)$. As this ratio is independent of the choice of π_Ψ, estimation of ψ is to a certain extent robust to the choice of prior. The role of the marginal prior π_Ψ arises in quantifying the uncertainty about the estimate of ψ through the regions C_γ. So the conditional prior given ψ, together with the model and data, are used to determine the form of any inferences about ψ,

while the marginal prior for ψ, together with the model and data, are used to quantify the uncertainty in these inferences.

By contrast predictions are based on the relative belief ratio $q(y \mid x)/q(y)$ which is generally dependent on the full prior π. So in a sense predictions are less robust to the prior than estimation. On the other hand Bayesian inferences are often advocated due to the regularizing effect of the prior. While the relative surprise approach does not fully incorporate such an effect for parameter estimates, the full effect is available for prediction.

Appendix

Proof of Theorem 2: We have that

$$\begin{aligned} r_\eta(\delta \mid x) &= \int_\Psi \frac{I(\psi \neq \delta(x))}{\max(\eta, \pi_\Psi(\psi))} \pi_\Psi(\psi \mid x)\, \nu_\Psi(d\psi) \\ &= \int_\Psi \frac{\pi_\Psi(\psi \mid x)}{\max(\eta, \pi_\Psi(\psi))}\, \nu_\Psi(d\psi) - \frac{\pi_\Psi(\delta(x) \mid x)}{\max(\eta, \pi_\Psi(\delta(x)))}. \end{aligned} \tag{17}$$

The first term in (17) is constant in $\delta(x)$ and bounded above by $1/\eta$, so the value of a Bayes rule at x is obtained by finding $\delta(x)$ that maximizes the second term.

Consider η as fixed and note that

$$\frac{\pi_\Psi(\delta(x) \mid x)}{\max(\eta, \pi_\Psi(\delta(x)))} = \begin{cases} \frac{\pi_\Psi(\delta(x) \mid x)}{\eta} & \text{if } \eta > \pi_\Psi(\delta(x)) \\ \frac{\pi_\Psi(\delta(x) \mid x)}{\pi_\Psi(\delta(x))} & \text{if } \eta \leq \pi_\Psi(\delta(x)). \end{cases} \tag{18}$$

There are at most finitely many values of ψ satisfying $\eta \leq \pi_\Psi(\psi)$ and so $\pi_\Psi(\psi \mid x)/\pi_\Psi(\psi)$ assumes a maximum on this set, say at $\psi_\eta(x)$. There are infinitely many values of ψ satisfying $\eta > \pi_\Psi(\psi)$ but clearly we can find $\eta' < \eta$ so that $\{\psi : \eta' < \pi_\Psi(\psi) < \eta\}$ is nonempty and finite. Thus, $\pi_\Psi(\psi \mid x)$ assumes its maximum on the set $\{\psi : \pi_\Psi(\psi) < \eta\}$ in the subset $\{\psi : \eta' < \pi_\Psi(\psi) < \eta\}$, say at $\psi'_\eta(x)$. Therefore, a Bayes rule $\delta_\eta(x)$ is given by $\delta_\eta(x) = \psi_\eta(x)$ when $\pi_\Psi(\psi_\eta(x) \mid x)/\pi_\Psi(\psi_\eta(x)) \geq \pi_\Psi(\psi'_\eta(x) \mid x)/\eta$ and $\delta_\eta(x) = \psi'_\eta(x)$ otherwise.

If $\eta > \pi_\Psi(\delta(x))$, then

$$\pi_\Psi(\delta(x) \mid x)/\eta < \pi_\Psi(\delta(x) \mid x)/\pi_\Psi(\delta(x)) \leq \pi_\Psi(\psi_{\mathrm{LRSE}}(x) \mid x)/\pi_\Psi(\psi_{\mathrm{LRSE}}(x)).$$

Therefore, whenever $\eta \leq \pi_\Psi(\psi_{\mathrm{LRSE}}(x))$ the maximizer of (18) is given by $\delta(x) = \psi_{\mathrm{LRSE}}(x)$ and the result is proved.

Proof of Theorem 6: Just as in Theorem 2 a Bayes rule $\delta_{\lambda,\eta}(x)$ maximizes $\pi_{\Psi,\lambda}(\delta(x) \mid x)/\max(\eta, \pi_{\Psi,\lambda}(\delta(x)))$ for $\delta(x) \in \Psi_\lambda$. Furthermore, as in Theorem 2, such a rule exists. Now define $\eta(\lambda)$ so that $0 < \eta(\lambda) < \Pi_\Psi(B_\lambda(\psi_{\mathrm{LRSE}}(x)))$. Note that $\eta(\lambda) \to 0$ as $\lambda \to 0$. We have that, as $\lambda \to 0$,

$$\begin{aligned} \frac{\pi_{\Psi,\lambda}(\psi_\lambda(\psi_{\mathrm{LRSE}}(x)) \mid x)}{\max(\eta(\lambda), \pi_{\Psi,\lambda}(\psi_\lambda(\psi_{\mathrm{LRSE}}(x))))} &= \frac{\pi_{\Psi,\lambda}(\psi_\lambda(\psi_{\mathrm{LRSE}}(x)) \mid x)}{\pi_{\Psi,\lambda}(\psi_\lambda(\psi_{\mathrm{LRSE}}(x)))} \\ &\to \frac{\pi_\Psi(\psi_{\mathrm{LRSE}}(x) \mid x)}{\pi_\Psi(\psi_{\mathrm{LRSE}}(x))}. \end{aligned} \tag{19}$$

Let $\epsilon > 0$. Let λ_0 be such that $\sup_{\psi\in\Psi}\mathrm{diam}(B_\lambda(\psi)) < \epsilon/2$ for all $\lambda < \lambda_0$. Then for $\lambda < \lambda_0$, and any $\delta(x)$ satisfying $||\delta(x) - \psi_{\mathrm{LRSE}}(x)|| \geq \epsilon$, we have

$$\begin{aligned}
\frac{\pi_{\Psi,\lambda}(\psi_\lambda(\delta(x)) \,|\, x)}{\pi_{\Psi,\lambda}(\psi_\lambda(\delta(x)))} &= \frac{\int_{B_\lambda(\psi_\lambda(\delta(x)))} \pi_\Psi(\psi \,|\, x)\, \nu_\Psi(d\psi)}{\int_{B_\lambda(\psi_\lambda(\delta(x)))} \pi_\Psi(\psi)\, \nu_\Psi(d\psi)} \\
&= \frac{\int_{B_\lambda(\psi_\lambda(\delta(x)))} \frac{\pi_\Psi(\psi \,|\, x)}{\pi_\Psi(\psi)} \pi_\Psi(\psi)\, \nu_\Psi(d\psi)}{\int_{B_\lambda(\psi_\lambda(\delta(x)))} \pi_\Psi(\psi)\, \nu_\Psi(d\psi)} \\
&\leq \sup_{\{\psi : ||\psi - \psi_{\mathrm{LRSE}}(x)|| > \epsilon/2\}} \frac{\pi_\Psi(\psi \,|\, x)}{\pi_\Psi(\psi)} < \frac{\pi_\Psi(\psi_{\mathrm{LRSE}}(x) \,|\, x)}{\pi_\Psi(\psi_{\mathrm{LRSE}}(x))}. \qquad (20)
\end{aligned}$$

By (19) and (20) there exists $\lambda_1 < \lambda_0$ such that, for all $\lambda < \lambda_1$,

$$\frac{\pi_{\Psi,\lambda}(\psi_\lambda(\psi_{\mathrm{LRSE}}(x)) \,|\, x)}{\pi_{\Psi,\lambda}(\psi_\lambda(\psi_{\mathrm{LRSE}}(x)))} > \sup_{\{\psi : ||\psi - \psi_{\mathrm{LRSE}}(x)|| > \epsilon/2\}} \frac{\pi_\Psi(\psi \,|\, x)}{\pi_\Psi(\psi)}. \qquad (21)$$

Therefore, when $\lambda < \lambda_1$, a Bayes rule $\delta_{\lambda,\eta(\lambda)}(x)$ satisfies

$$\begin{aligned}
\frac{\pi_{\Psi,\lambda}(\delta_{\lambda,\eta(\lambda)}(x) \,|\, x)}{\pi_{\Psi,\lambda}(\delta_{\lambda,\eta(\lambda)}(x))} &\geq \frac{\pi_{\Psi,\lambda}(\delta_{\lambda,\eta(\lambda)}(x) \,|\, x)}{\max(\eta(\lambda), \pi_{\Psi,\lambda}(\delta_{\lambda,\eta(\lambda)}(x)))} \\
&\geq \frac{\pi_{\Psi,\lambda}(\psi_\lambda(\psi_{\mathrm{LRSE}}(x)) \,|\, x)}{\max(\eta(\lambda), \pi_{\Psi,\lambda}(\psi_\lambda(\psi_{\mathrm{LRSE}}(x))))} = \frac{\pi_{\Psi,\lambda}(\psi_\lambda(\psi_{\mathrm{LRSE}}(x)) \,|\, x)}{\pi_{\Psi,\lambda}(\psi_\lambda(\psi_{\mathrm{LRSE}}(x)))}. \qquad (22)
\end{aligned}$$

By (20), (21) and (22) this implies that $||\delta_{\lambda,\eta(\lambda)} - \psi_{\mathrm{LRSE}}(x)|| < \epsilon$ and the convergence is established.

Proof of Corollary 7: Following the proof of Theorem 6 we have that $\pi_{\Psi,\lambda}(\hat{\psi}_\lambda(x) \,|\, x)/\pi_{\Psi,\lambda}(\hat{\psi}_\lambda(x)) \geq \pi_{\Psi,\lambda}(\delta_{\lambda,\eta(\lambda)}(x) \,|\, x)/\pi_{\Psi,\lambda}(\delta_{\lambda,\eta(\lambda)}(x))$ and so by (20), (21) and (22) this implies that $||\hat{\psi}_\lambda(x) - \psi_{\mathrm{LRSE}}(x)|| < \epsilon$ and the convergence of $\hat{\psi}_\lambda(x)$ to $\psi_{\mathrm{LRSE}}(x)$ is established.

Proof of Theorem 9: For $c > 0$ let $S_c(x) = \{\pi_\Psi(\psi \,|\, x)/\pi_\Psi(\psi) \geq c\}$ and $S_{\eta,c}(x) = \{\pi_\Psi(\psi \,|\, x)/\max(\eta, \pi_\Psi(\psi)) \geq c\}$. Note that $S_{\eta,c}(x) \uparrow S_c(x)$ as $\eta \to 0$.

Suppose c is such that $\Pi_\Psi(S_c(x) \,|\, x) \leq \gamma$. Then $\Pi_\Psi(S_{\eta,c}(x) \,|\, x) \leq \gamma$ for all η and so $S_{\eta,c}(x) \subset L_{\eta,\gamma}(x)$. This implies that $S_c(x) \subset \liminf_{\eta\to 0} L_{\eta,\gamma}(x)$ and since $\Pi_\Psi(C_\gamma(x) \,|\, x) = \gamma$ this implies that $C_\gamma(x) \subset \liminf_{\eta\to 0} L_{\eta,\gamma}(x)$.

Now suppose c is such that $\Pi_\Psi(S_c(x) \,|\, x) > \gamma$. Then there exists η_0 such that for all $\eta < \eta_0$ we have $\Pi_\Psi(S_{\eta,c}(x) \,|\, x) > \gamma$. Since $L_{\eta,\gamma}(x) \subset S_{\eta,c}(x)$ we have that $\limsup_{\eta\to 0} L_{\eta,\gamma}(x) \subset S_c(x)$. Then choosing $c = c_{\gamma'}(x)$ for $\gamma' > \gamma$ implies that $\limsup_{\eta\to 0} L_{\eta,\gamma}(x) \subset C_{\gamma'}(x)$.

Proof of Theorem 10: Let $S_c(x) = \{\psi : \pi_\Psi(\psi \,|\, x)/\pi_\Psi(\psi) \geq c\}$ and $S_{\lambda,c}(x) = \{\psi : \Pi_\Psi(B_\lambda(\psi) \,|\, x)/\Pi_\Psi(B_\lambda(\psi)) \geq c\}$. Recall that

$$\lim_{\lambda\to 0} \Pi_\Psi(B_\lambda(\psi) \,|\, x)/\Pi_\Psi(B_\lambda(\psi)) = \pi_\Psi(\psi \,|\, x)/\pi_\Psi(\psi)$$

for every ψ. If $\pi_\Psi(\psi \,|\, x)/\pi_\Psi(\psi) > c$, we have that there exists λ_0 such that for all $\lambda < \lambda_0$, then $\Pi_\Psi(B_\lambda(\psi) \,|\, x)/\Pi_\Psi(B_\lambda(\psi)) > c$ and this implies that $\psi \in$

$\liminf_{\lambda\to 0} S_{\lambda,c}(x)$. Now $\Pi_\Psi(\pi_\Psi(\psi\,|\,x)/\pi_\Psi(\psi) = c) = 0$ and so we have $S_c(x) \subset \liminf_{\lambda\to 0} S_{\lambda,c}(x)$ (after possibly deleting a set of Π_Ψ-measure 0 from $S_c(x)$). Now, if $\psi \in \limsup_{\lambda\to 0} S_{\lambda,c}(x)$, then $\Pi_\Psi(B_\lambda(\psi)\,|\,x)/\Pi_\Psi(B_\lambda(\psi))$
$\geq c$ for infinitely many $\lambda \to 0$, which implies that $\pi_\Psi(\psi\,|\,x)/\pi_\Psi(\psi) \geq c$, and therefore $\psi \in S_c(x)$. This proves $S_c(x) = \lim_{\lambda\to 0} S_{\lambda,c}(x)$ (up to a set of Π_Ψ-measure 0) so that $\lim_{\lambda\to 0} \Pi_\Psi(S_{\lambda,c}(x)\Delta S_c(x)\,|\,x) = 0$ for any c.

Let $c_{\lambda,\gamma}(x) = \sup\{c \geq 0 : \Pi_\Psi(S_{\lambda,c}(x)\,|\,x) \geq \gamma\}$ so $S_{c_\gamma(x)}(x) = C_\gamma(x)$ and $S_{\lambda,c_{\lambda,\gamma}(x)}(x) = C_{\lambda,\gamma}(x)$. Then we have that

$$\begin{aligned}\Pi_\Psi(C_\gamma(x)\Delta C_{\lambda,\gamma}(x)\,|\,x) &= \Pi_\Psi(S_{c_\gamma(x)}(x)\Delta S_{\lambda,c_{\lambda,\gamma}(x)}(x)\,|\,x)\\ &\leq \Pi_\Psi(S_{c_\gamma(x)}(x)\Delta S_{\lambda,c_\gamma(x)}(x)\,|\,x) + \Pi_\Psi(S_{\lambda,c_{\lambda,\gamma}(x)}(x)\Delta S_{\lambda,c_\gamma(x)}(x)\,|\,x). \qquad (23)\end{aligned}$$

Since $S_{c_\gamma(x)}(x) = \lim_{\lambda\to 0} S_{\lambda,c_\gamma(x)}(x)$ we have $\Pi_\Psi(S_{c_\gamma(x)}(x)\Delta S_{\lambda,c_\gamma(x)}(x)\,|\,x) \to 0$ and $\Pi_\Psi(S_{\lambda,c_\gamma(x)}(x)\,|\,x) \to \Pi_\Psi(S_{c_\gamma(x)}(x)\,|\,x) = \gamma$ as $\lambda \to 0$. Now consider the second term in (23). Since $\pi_\Psi(\psi\,|\,x)/\pi_\Psi(\psi)$ has a continuous posterior distribution, we have $\Pi_\Psi(\pi_\Psi(\psi\,|\,x)/\pi_\Psi(\psi) \geq c\,|\,x)$ is continuous in c. Let $\epsilon > 0$ and note that for all λ small enough, $\Pi_\Psi(S_{\lambda,c_{\gamma-\epsilon}(x)}(x)\,|\,x) < \gamma$ and $\Pi_\Psi(S_{\lambda,c_{\gamma+\epsilon}(x)}(x)\,|\,x) > \gamma$ which implies that $c_{\gamma+\epsilon}(x) \leq c_{\lambda,\gamma}(x) \leq c_{\gamma-\epsilon}(x)$ and therefore $S_{\lambda,c_{\gamma+\epsilon}(x)}(x) \subset S_{\lambda,c_{\lambda,\gamma}(x)} \subset S_{\lambda,c_{\gamma-\epsilon}(x)}(x)$. As $S_{\lambda,c_{\lambda,\gamma}(x)}(x) \subset S_{\lambda,c_\gamma(x)}(x)$ or $S_{\lambda,c_{\lambda,\gamma}(x)}(x) \supset S_{\lambda,c_\gamma(x)}(x)$ then

$$\Pi_\Psi(S_{\lambda,c_{\lambda,\gamma}(x)}(x)\Delta S_{\lambda,c_\gamma(x)}(x)\,|\,x) = |\Pi_\Psi(S_{\lambda,c_{\lambda,\gamma}(x)}(x)\,|\,x) - \Pi_\Psi(S_{\lambda,c_\gamma(x)}(x)\,|\,x)|.$$

For all λ small $|\Pi_\Psi(S_{\lambda,c_{\lambda,\gamma}(x)}(x)\,|\,x) - \Pi_\Psi(S_{\lambda,c_\gamma(x)}(x)\,|\,x)|$ is bounded above by

$$\begin{aligned}\max\{&|\Pi_\Psi(S_{\lambda,c_{\gamma+\epsilon}(x)}(x)\,|\,x) - \Pi_\Psi(S_{\lambda,c_\gamma(x)}(x)\,|\,x)|,\\ &|\Pi_\Psi(S_{\lambda,c_{\gamma-\epsilon}(x)}(x)\,|\,x) - \Pi_\Psi(S_{\lambda,c_\gamma(x)}(x)\,|\,x)|\}\end{aligned}$$

and this upper bound converges to ϵ as $\lambda \to 0$. Since ϵ is arbitrary we have that the second term in (23) goes to 0 as $\lambda \to 0$ and this proves the result.

Proof of Theorem 11: Suppose, without loss of generality that $0 < \gamma < 1$. Let $\epsilon > 0$ and $\delta > 0$ satisfy $\gamma + \delta \leq 1$. Put $\gamma'(\lambda,\gamma) = \Pi_\Psi(C_{\lambda,\gamma}(x)\,|\,x), \gamma''(\lambda,\gamma)$
$= \Pi_\Psi(C_{\gamma+\delta}(x)\,|\,x)$ and note that $\gamma'(\lambda,\gamma) \geq \gamma, \gamma''(\lambda,\gamma) \geq \gamma + \delta$. By Theorem 10 we have that $C_{\lambda,\gamma}(x) \to C_\gamma(x)$ and $C_{\lambda,\gamma+\delta}(x) \to C_{\gamma+\delta}(x)$ as $\lambda \to 0$ so $\gamma'(\lambda,\gamma) \to \gamma$ and $\gamma''(\lambda,\gamma) \to \gamma + \delta$ as $\lambda \to 0$. This implies that there is a $\lambda_0(\delta)$ such that for all $\lambda < \lambda_0(\delta)$ then $\gamma'(\lambda,\gamma) < \gamma''(\lambda,\gamma)$. Therefore, by Theorem 9, we have that for all $\lambda < \lambda_0(\delta)$

$$C_{\lambda,\gamma}(x) \subset \liminf_{\eta\to 0} L_{\eta,\lambda,\gamma'(\lambda,\gamma)}(x) \subset \limsup_{\eta\to 0} L_{\eta,\lambda,\gamma'(\lambda,\gamma)}(x) \subset C_{\lambda,\gamma+\delta}(x). \qquad (24)$$

From (24) and Theorem 10 we have that $C_\gamma(x) \subset \liminf_{\lambda\to 0} \liminf_{\eta\to 0} L_{\eta,\lambda,\gamma'(\lambda,\gamma)}(x)$
$\subset \limsup_{\lambda\to 0} \limsup_{\eta\to 0} L_{\eta,\lambda,\gamma'(\lambda,\gamma)}(x) \subset C_{\gamma+\delta}(x)$. Since $\lim_{\delta\to 0} C_{\gamma+\delta}(x) = C_\gamma(x)$ this establishes the result.

References

Aitkin, M. (2010). Statistical Inference: An Integrated Bayesian/Likelihood Approach. Chapman and Hall/CRC, Boca Raron, FL.

Baldi, P. and Itti, L. (2010). Of bits and wows: A Bayesian theory of surprise with applications to attention. Neural Networks, 23, 649-666.

Berger, J.O. (1985). Statistical Decision Theory and Bayesian Analysis. Springer, New York.

Berger, J.O., Liseo, B. and Wolpert, R.L. (1999). Integrated likelihood methods for eliminating nuisance parameters. Stat. Sci., 14(1):1-28.

Bernardo, J. M. (2005). Intrinsic credible regions: an objective Bayesian approach to interval estimation. Test, 14(2):317–384. With comments and a rejoinder by the author.

Bernardo, J. M. and Smith, A. F. M. (2000). Bayesian Theory. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., New York. Paperback.

Bishop, C.M. (2006). Pattern Recognition and Machine Learning. Springer, New York.

Dempster, A. P. (1973). The direct use of likelihood for significance testing. Memoirs, No. 1, Proceedings of Conference on Foundational Questions in Statistical Inference, eds. O. Barndorff-Nielsen, P. Blaesild and G. Schou, Institute of Mathematics, U. of Aarhus, 335-354 (reprinted in Statistics and Computing (1997), 7, 247-252).

Evans, M. (1997). Bayesian inference procedures derived via the concept of relative surprise. Comm. Statist. Theory Methods, 26(5):1125–1143.

Evans, M. J., Guttman, I., and Swartz, T. (2006). Optimality and computations for relative surprise inferences. Canad. J. Statist, 34(1):113-129.

Evans, M. and Shakhatreh, M. (2008). Optimal properties of some Bayesian inferences. Electron. J. Stat., 2, 1268–1280.

Jang, G. H. (2010). Invariant Procedures for Model Checking, Checking for Prior-Data Conflict and Bayesian Inference. Ph.D. thesis, University of Toronto.

Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. Univ. California Publ. Statist, 1, 277–329.

Robert, C.P. (1996). Intrinsic losses. Theory and Decision, 40, 191-214.

Rudin, W. (1974). Real and Complex Analysis. McGraw Hill, New York.

Tjur, T. (1974). Conditional Probability Distributions. Institute of Mathematical Statistics, University of Copenhagen, Copenhagen.